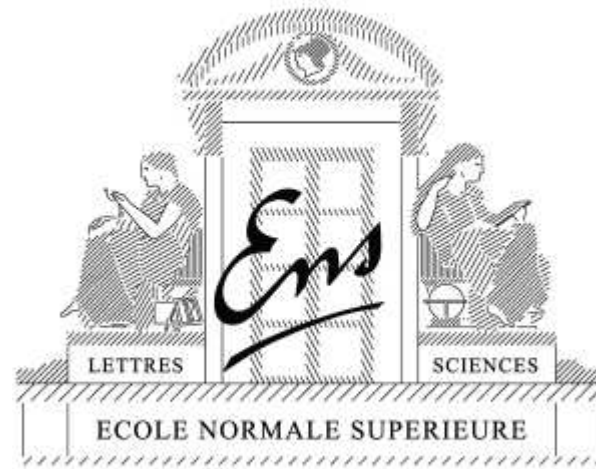


Beyond stochastic gradient descent for large-scale machine learning

Francis Bach

INRIA - Ecole Normale Supérieure, Paris, France



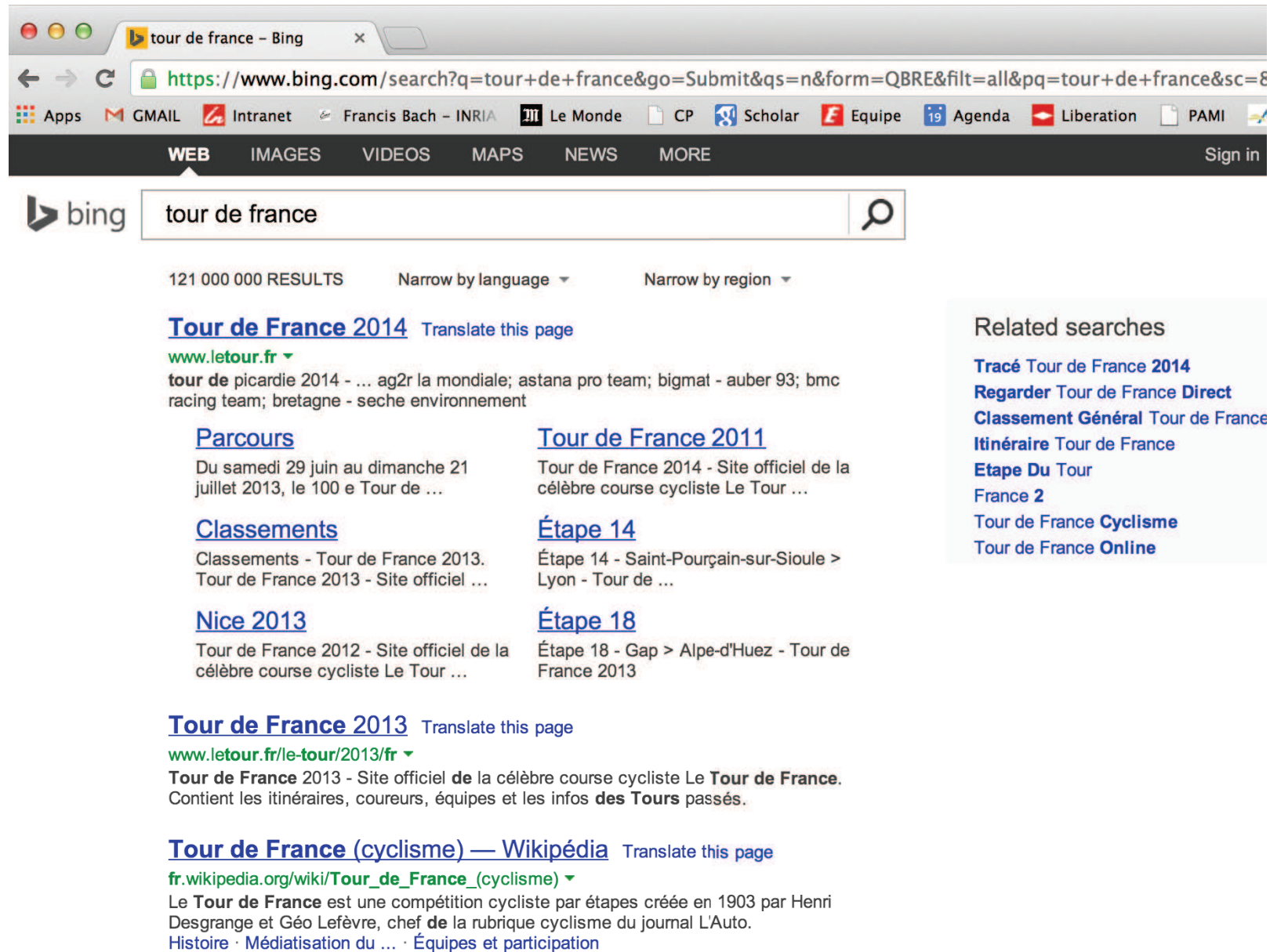
Joint work with Eric Moulines, Nicolas Le Roux
and Mark Schmidt - CAP, July 2014

“Big data” revolution?

A new scientific context

- **Data everywhere:** size does not (always) matter
- **Science and industry**
- **Size and variety**
- **Learning from examples**
 - n observations in dimension p

Search engines - Advertising



The image shows a screenshot of a web browser displaying a Bing search results page for the query "tour de france". The browser's address bar shows the URL: <https://www.bing.com/search?q=tour+de+france&go=Submit&qsn=n&form=QBRE&filt=all&pq=tour+de+france&sc=8>. The search bar contains the text "tour de france". Below the search bar, the results show "121 000 000 RESULTS" and options to "Narrow by language" and "Narrow by region".

The first search result is for "Tour de France 2014", with a link to www.letour.fr. The snippet reads: "tour de picardie 2014 - ... ag2r la mondiale; astana pro team; bigmat - auber 93; bmc racing team; bretagne - seche environnement".

Below this result are several sub-links:

- [Parcours](#): Du samedi 29 juin au dimanche 21 juillet 2013, le 100 e Tour de ...
- [Classements](#): Classements - Tour de France 2013. Tour de France 2013 - Site officiel ...
- [Nice 2013](#): Tour de France 2012 - Site officiel de la célèbre course cycliste Le Tour ...
- [Tour de France 2011](#): Tour de France 2014 - Site officiel de la célèbre course cycliste Le Tour ...
- [Étape 14](#): Étape 14 - Saint-Pourçain-sur-Sioule > Lyon - Tour de ...
- [Étape 18](#): Étape 18 - Gap > Alpe-d'Huez - Tour de France 2013

On the right side of the page, there is a "Related searches" section with the following links:

- [Tracé Tour de France 2014](#)
- [Regarder Tour de France Direct](#)
- [Classement Général Tour de France](#)
- [Itinéraire Tour de France](#)
- [Étape Du Tour France 2](#)
- [Tour de France Cyclisme](#)
- [Tour de France Online](#)

The second search result is for "Tour de France 2013", with a link to www.letour.fr/le-tour/2013/fr. The snippet reads: "Tour de France 2013 - Site officiel de la célèbre course cycliste Le Tour de France. Contient les itinéraires, coureurs, équipes et les infos des Tours passés."

The third search result is for "Tour de France (cyclisme) — Wikipédia", with a link to [fr.wikipedia.org/wiki/Tour_de_France_\(cyclisme\)](http://fr.wikipedia.org/wiki/Tour_de_France_(cyclisme)). The snippet reads: "Le Tour de France est une compétition cycliste par étapes créée en 1903 par Henri Desgrange et Géo Lefèvre, chef de la rubrique cyclisme du journal L'Auto. Histoire · Médiatisation du ... · Équipes et participation".

Marketing - Personalized recommendation

Amazon.com: Online Shopping | Google Search

www.amazon.com

Le Monde | Intranet INRIA | Francis Bach | GMAIL | Liberation | L'EQUIPE | Google Scholar | PAMI | iGoogle | CP | StatCounter | Analytics | Zimbra

amazon

FRANCIS's Amazon.com | Today's Deals | Gift Cards | Help

The All-New kindle fire HD

Shop by Department | Search All | Go

Hello, FRANCIS Your Account | Cart | Wish List

Achetez-vous depuis la France? Shopping from France? Essayez [amazon.fr](#) > Cliquez ici

amazon Get the Free Amazon Mobile App Search & buy millions of products on the go > Learn more

Instant Video | MP3 Store | Cloud Player | **Kindle** | Cloud Drive | Appstore for Android | Digital Games & Software | Audible Audiobooks

The All-New Kindle Family

Kindle Paperwhite \$119
Kindle Fire HD \$199
Kindle Fire HD 8.9" \$299



Bikes with Street Cred | **Clothing Trends** | Amazon Prime

THE AMAZON CLOTHING STORE

Color Theory

Bright outerwear by Nicole Miller, Calvin Klein, Diesel, and more.

> View Looks
> Shop All Clothing

Understand what the **Zeros and Ones** are telling you.

Learn more

Advertisement

3M Streaming Projector Powered by Roku

Pre-order now for \$20 Amazon Instant Video credit > Learn more

Personal photos


photos ete 2009

FAVORITES


- Dropbox
- BOOKS
- Applications
- fback
- Desktop
- Downloads

SHARED

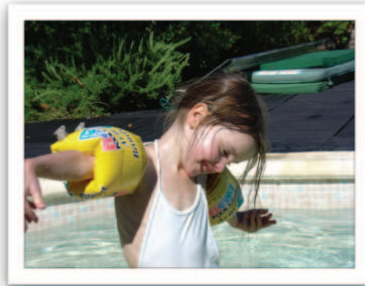
TAGS




DSC07338.JPG




DSC07343.JPG




DSC07344.JPG



DSC07348.JPG

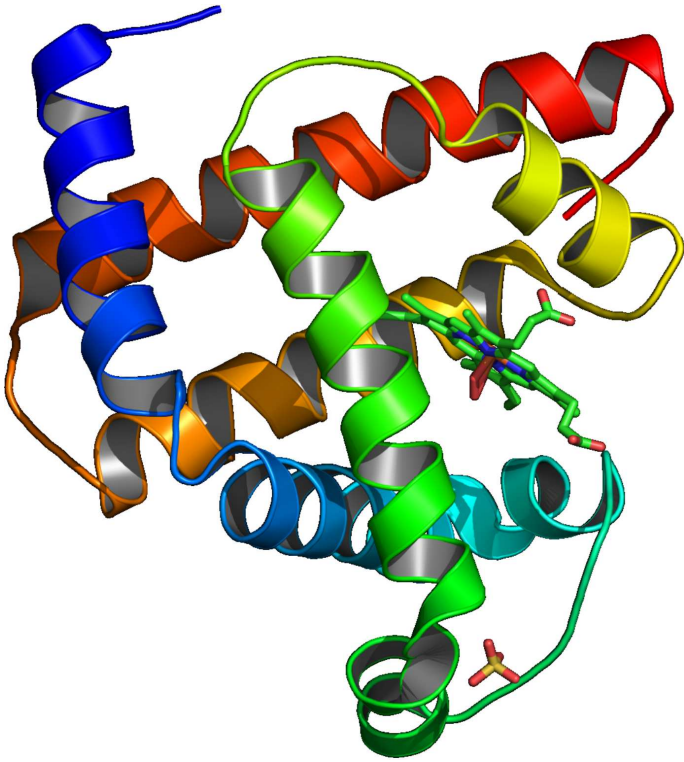


DSC07349.JPG



DSC07350.JPG

Bioinformatics



- **Protein:** Crucial elements of cell life
- **Massive data:** 2 millions for humans
- **Complex data**

Context

Machine learning for “big data”

- **Large-scale machine learning:** **large p , large n**
 - p : dimension of each observation (input)
 - n : number of observations
- **Examples:** computer vision, bioinformatics, advertising

Context

Machine learning for “big data”

- **Large-scale machine learning:** **large p , large n**
 - p : dimension of each observation (input)
 - n : number of observations
- **Examples:** computer vision, bioinformatics, advertising
- **Ideal running-time complexity:** $O(pn)$

Context

Machine learning for “big data”

- **Large-scale machine learning:** **large p , large n**
 - p : dimension of each observation (input)
 - n : number of observations
- **Examples:** computer vision, bioinformatics, advertising
- **Ideal running-time complexity:** $O(pn)$
- **Going back to simple methods**
 - Stochastic gradient methods (Robbins and Monro, 1951)
 - Mixing statistics and optimization

Outline

- **Introduction: stochastic approximation algorithms**
 - Supervised machine learning and convex optimization
 - Stochastic gradient and averaging
 - Strongly convex vs. non-strongly convex
- **Fast convergence through smoothness and constant step-sizes**
 - Online Newton steps (Bach and Moulines, 2013)
 - $O(1/n)$ convergence rate for all convex functions
- **More than a single pass through the data**
 - Stochastic average gradient (Le Roux, Schmidt, and Bach, 2012)
 - Linear (exponential) convergence rate for strongly convex functions

Supervised machine learning

- **Data:** n observations $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \dots, n$, **i.i.d.**
- Prediction as a linear function $\langle \theta, \Phi(x) \rangle$ of features $\Phi(x) \in \mathbb{R}^p$
- **(regularized) empirical risk minimization:** find $\hat{\theta}$ solution of

$$\min_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \ell(y_i, \langle \theta, \Phi(x_i) \rangle) + \mu \Omega(\theta)$$

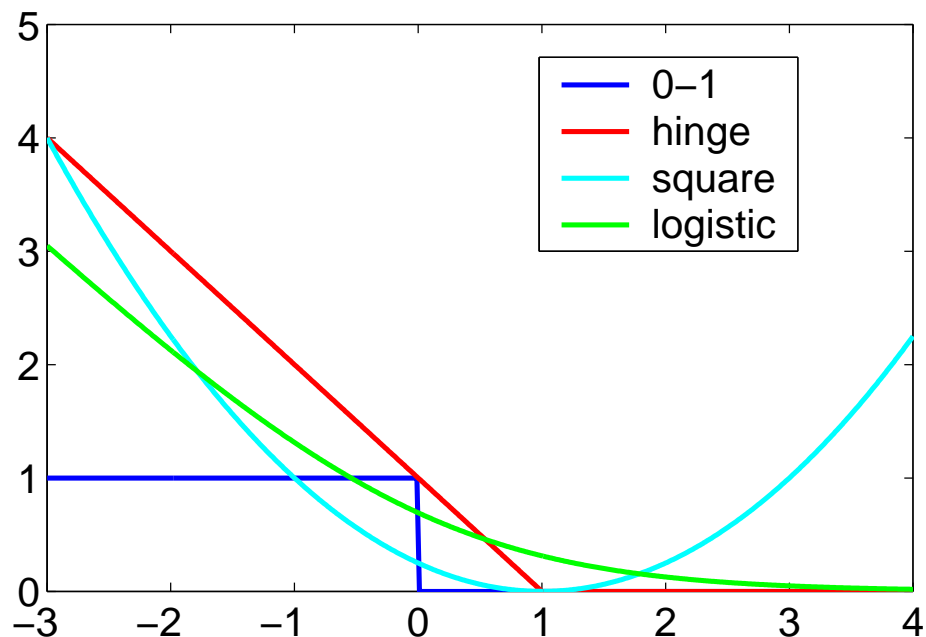
convex data fitting term + regularizer

Usual losses

- **Regression:** $y \in \mathbb{R}$, prediction $\hat{y} = \langle \theta, \Phi(x) \rangle$
 - quadratic loss $\frac{1}{2}(y - \hat{y})^2 = \frac{1}{2}(y - \langle \theta, \Phi(x) \rangle)^2$

Usual losses

- **Regression:** $y \in \mathbb{R}$, prediction $\hat{y} = \langle \theta, \Phi(x) \rangle$
 - quadratic loss $\frac{1}{2}(y - \hat{y})^2 = \frac{1}{2}(y - \langle \theta, \Phi(x) \rangle)^2$
- **Classification :** $y \in \{-1, 1\}$, prediction $\hat{y} = \text{sign}(\langle \theta, \Phi(x) \rangle)$
 - loss of the form $\ell(y \langle \theta, \Phi(x) \rangle)$
 - “True” **0-1** loss: $\ell(y \langle \theta, \Phi(x) \rangle) = 1_{y \langle \theta, \Phi(x) \rangle < 0}$
 - Usual **convex** losses:



Supervised machine learning

- **Data:** n observations $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \dots, n$, **i.i.d.**
- Prediction as a linear function $\langle \theta, \Phi(x) \rangle$ of features $\Phi(x) \in \mathbb{R}^p$
- **(regularized) empirical risk minimization:** find $\hat{\theta}$ solution of

$$\min_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \ell(y_i, \langle \theta, \Phi(x_i) \rangle) + \mu \Omega(\theta)$$

convex data fitting term + regularizer

Supervised machine learning

- **Data:** n observations $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \dots, n$, **i.i.d.**
- Prediction as a linear function $\langle \theta, \Phi(x) \rangle$ of features $\Phi(x) \in \mathbb{R}^p$
- **(regularized) empirical risk minimization:** find $\hat{\theta}$ solution of

$$\min_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \ell(y_i, \langle \theta, \Phi(x_i) \rangle) + \mu \Omega(\theta)$$

convex data fitting term + regularizer

- Empirical risk: $\hat{f}(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, \langle \theta, \Phi(x_i) \rangle)$ **training cost**
- Expected risk: $f(\theta) = \mathbb{E}_{(x,y)} \ell(y, \langle \theta, \Phi(x) \rangle)$ **testing cost**
- **Two fundamental questions:** (1) computing $\hat{\theta}$ and (2) analyzing $\hat{\theta}$

Supervised machine learning

- **Data:** n observations $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \dots, n$, **i.i.d.**
- Prediction as a linear function $\langle \theta, \Phi(x) \rangle$ of features $\Phi(x) \in \mathbb{R}^p$
- **(regularized) empirical risk minimization:** find $\hat{\theta}$ solution of

$$\min_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \ell(y_i, \langle \theta, \Phi(x_i) \rangle) + \mu \Omega(\theta)$$

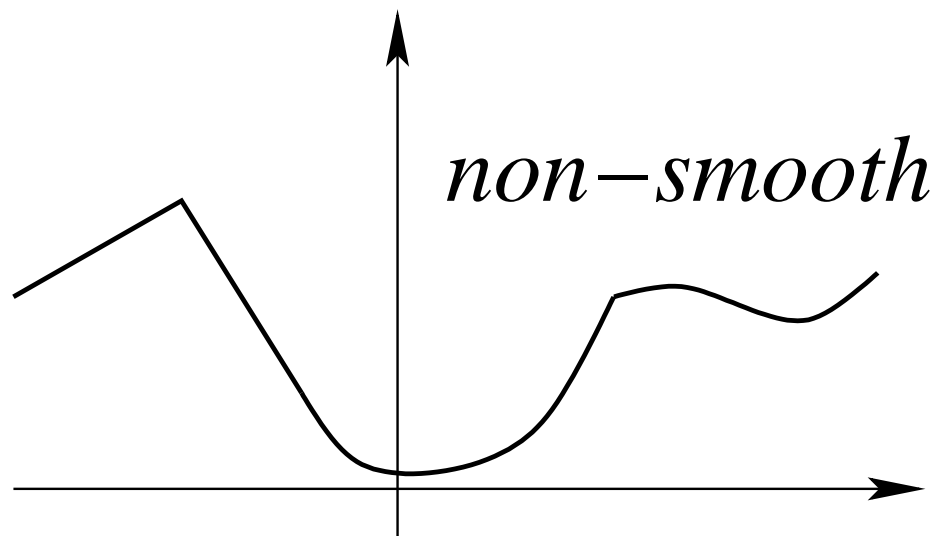
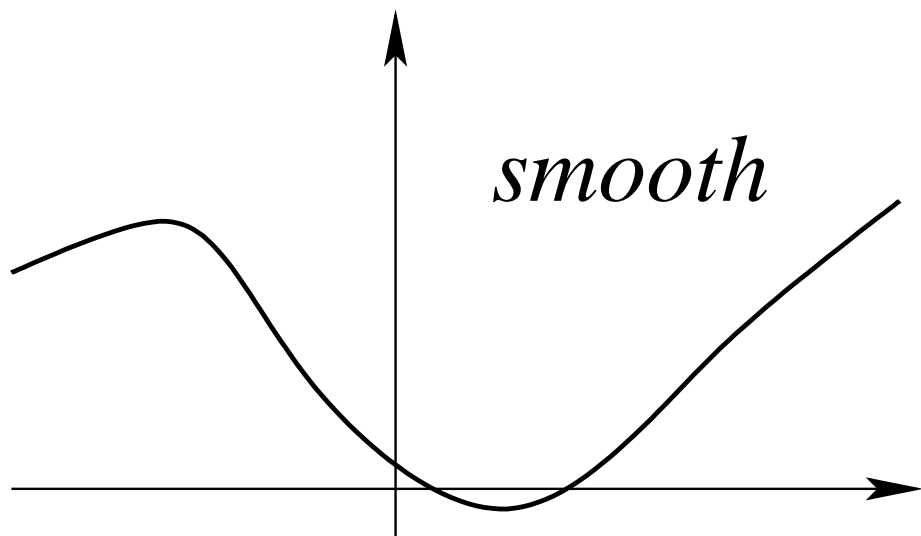
convex data fitting term + regularizer

- Empirical risk: $\hat{f}(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, \langle \theta, \Phi(x_i) \rangle)$ **training cost**
- Expected risk: $f(\theta) = \mathbb{E}_{(x,y)} \ell(y, \langle \theta, \Phi(x) \rangle)$ **testing cost**
- **Two fundamental questions:** (1) computing $\hat{\theta}$ and (2) analyzing $\hat{\theta}$
 - **May be tackled simultaneously**

Smoothness and strong convexity

- A function $g : \mathbb{R}^p \rightarrow \mathbb{R}$ is L -smooth if and only if it is twice differentiable and

$$\forall \theta \in \mathbb{R}^p, g''(\theta) \preceq L \cdot \text{Id}$$



Smoothness and strong convexity

- A function $g : \mathbb{R}^p \rightarrow \mathbb{R}$ is **L -smooth** if and only if it is twice differentiable and

$$\forall \theta \in \mathbb{R}^p, g''(\theta) \preceq L \cdot \text{Id}$$

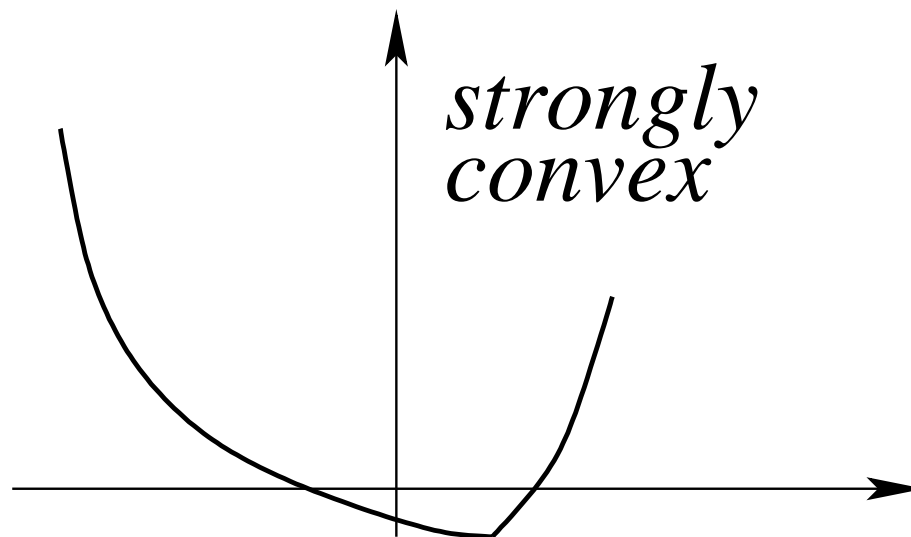
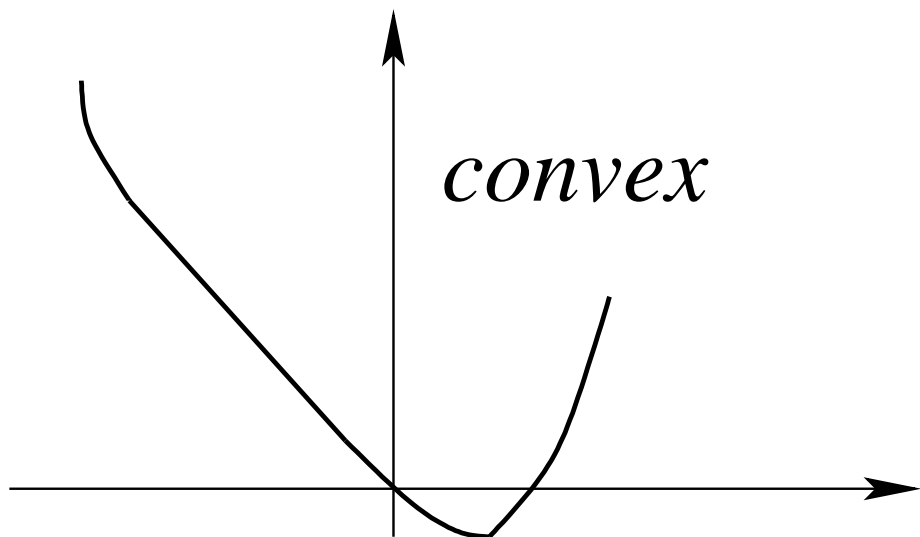
- **Machine learning**

- with $g(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, \langle \theta, \Phi(x_i) \rangle)$
- Hessian \approx covariance matrix $\frac{1}{n} \sum_{i=1}^n \Phi(x_i) \otimes \Phi(x_i)$
- **Bounded data**

Smoothness and strong convexity

- A twice differentiable function $g : \mathbb{R}^p \rightarrow \mathbb{R}$ is μ -strongly convex if and only if

$$\forall \theta \in \mathbb{R}^p, g''(\theta) \succcurlyeq \mu \cdot \text{Id}$$



Smoothness and strong convexity

- A twice differentiable function $g : \mathbb{R}^p \rightarrow \mathbb{R}$ is μ -strongly convex if and only if

$$\forall \theta \in \mathbb{R}^p, g''(\theta) \succcurlyeq \mu \cdot \text{Id}$$

- **Machine learning**

- with $g(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, \langle \theta, \Phi(x_i) \rangle)$
- Hessian \approx covariance matrix $\frac{1}{n} \sum_{i=1}^n \Phi(x_i) \otimes \Phi(x_i)$
- **Data with invertible covariance matrix** (low correlation/dimension)

Smoothness and strong convexity

- A twice differentiable function $g : \mathbb{R}^p \rightarrow \mathbb{R}$ is μ -strongly convex if and only if

$$\forall \theta \in \mathbb{R}^p, g''(\theta) \succcurlyeq \mu \cdot \text{Id}$$

- **Machine learning**

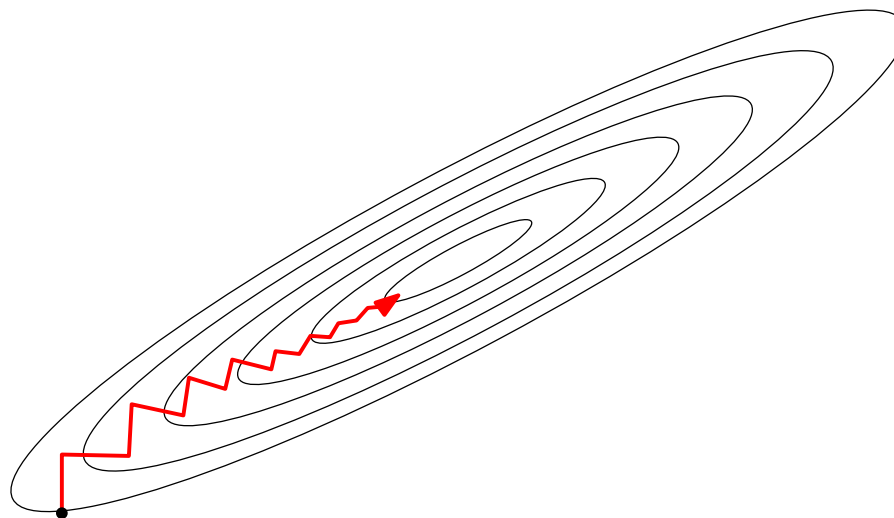
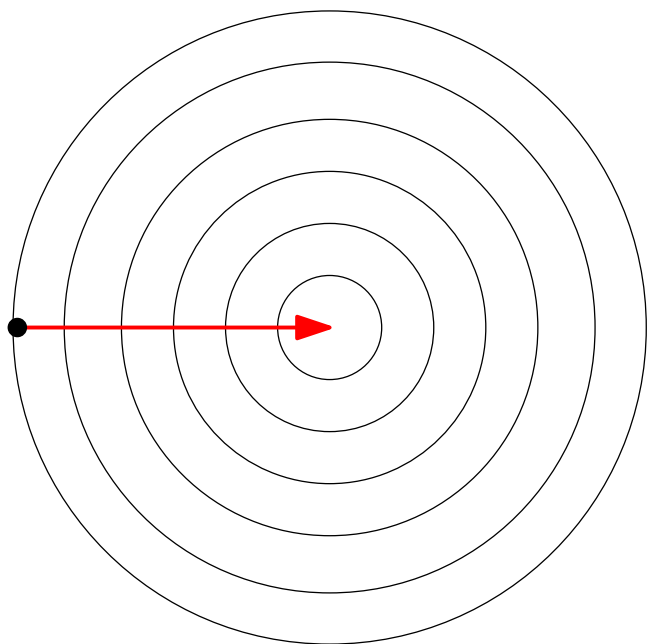
- with $g(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, \langle \theta, \Phi(x_i) \rangle)$
- Hessian \approx covariance matrix $\frac{1}{n} \sum_{i=1}^n \Phi(x_i) \otimes \Phi(x_i)$
- **Data with invertible covariance matrix** (low correlation/dimension)

- **Adding regularization by $\frac{\mu}{2} \|\theta\|^2$**

- **creates additional bias unless μ is small**

Iterative methods for minimizing smooth functions

- **Assumption:** g convex and smooth on \mathbb{R}^p
- **Gradient descent:** $\theta_t = \theta_{t-1} - \gamma_t g'(\theta_{t-1})$
 - $O(1/t)$ convergence rate for convex functions
 - $O(e^{-\rho t})$ convergence rate for strongly convex functions



Iterative methods for minimizing smooth functions

- **Assumption:** g convex and smooth on \mathbb{R}^p
- **Gradient descent:** $\theta_t = \theta_{t-1} - \gamma_t g'(\theta_{t-1})$
 - $O(1/t)$ convergence rate for convex functions
 - $O(e^{-\rho t})$ convergence rate for strongly convex functions
- **Newton method:** $\theta_t = \theta_{t-1} - g''(\theta_{t-1})^{-1} g'(\theta_{t-1})$
 - $O(e^{-\rho 2^t})$ convergence rate

Iterative methods for minimizing smooth functions

- **Assumption:** g convex and smooth on \mathbb{R}^p
- **Gradient descent:** $\theta_t = \theta_{t-1} - \gamma_t g'(\theta_{t-1})$
 - $O(1/t)$ convergence rate for convex functions
 - $O(e^{-\rho t})$ convergence rate for strongly convex functions
- **Newton method:** $\theta_t = \theta_{t-1} - g''(\theta_{t-1})^{-1} g'(\theta_{t-1})$
 - $O(e^{-\rho 2^t})$ convergence rate
- **Key insights from Bottou and Bousquet (2008)**
 1. In machine learning, no need to optimize below statistical error
 2. In machine learning, cost functions are averages

\Rightarrow **Stochastic approximation**

Stochastic approximation

- **Goal:** Minimizing a function f defined on \mathbb{R}^p
 - given only unbiased estimates $f'_n(\theta_n)$ of its gradients $f'(\theta_n)$ at certain points $\theta_n \in \mathbb{R}^p$

Stochastic approximation

- **Goal:** Minimizing a function f defined on \mathbb{R}^p
 - given only unbiased estimates $f'_n(\theta_n)$ of its gradients $f'(\theta_n)$ at certain points $\theta_n \in \mathbb{R}^p$
- **Machine learning - statistics**
 - $f(\theta) = \mathbb{E} f_n(\theta) = \mathbb{E} \ell(y_n, \langle \theta, \Phi(x_n) \rangle) =$ **generalization error**
 - **Loss for a single pair of observations:** $f_n(\theta) = \ell(y_n, \langle \theta, \Phi(x_n) \rangle)$
 - Expected gradient:

$$f'(\theta) = \mathbb{E} f'_n(\theta) = \mathbb{E} \{ \ell'(y_n, \langle \theta, \Phi(x_n) \rangle) \Phi(x_n) \}$$

- Beyond convex optimization: see, e.g., Benveniste et al. (2012)

Convex stochastic approximation

- **Key assumption:** smoothness and/or strong convexity
- **Key algorithm:** stochastic gradient descent (a.k.a. Robbins-Monro)

$$\theta_n = \theta_{n-1} - \gamma_n f'_n(\theta_{n-1})$$

– Polyak-Ruppert averaging: $\bar{\theta}_n = \frac{1}{n+1} \sum_{k=0}^n \theta_k$

– Which learning rate sequence γ_n ? Classical setting:

$$\gamma_n = Cn^{-\alpha}$$

Convex stochastic approximation

- **Key assumption:** smoothness and/or strong convexity
- **Key algorithm:** stochastic gradient descent (a.k.a. Robbins-Monro)

$$\theta_n = \theta_{n-1} - \gamma_n f'_n(\theta_{n-1})$$

– Polyak-Ruppert averaging: $\bar{\theta}_n = \frac{1}{n+1} \sum_{k=0}^n \theta_k$

– Which learning rate sequence γ_n ? Classical setting:

$$\gamma_n = Cn^{-\alpha}$$

- **Running-time** = $O(np)$
 - Single pass through the data
 - One line of code among many

Convex stochastic approximation

Existing work

- Known **global** minimax rates of convergence for **non-smooth** problems (Nemirovsky and Yudin, 1983; Agarwal et al., 2012)
 - **Strongly convex:** $O((\mu n)^{-1})$
Attained by averaged stochastic gradient descent with $\gamma_n \propto (\mu n)^{-1}$
 - **Non-strongly convex:** $O(n^{-1/2})$
Attained by averaged stochastic gradient descent with $\gamma_n \propto n^{-1/2}$

Convex stochastic approximation

Existing work

- Known **global** minimax rates of convergence for **non-smooth** problems (Nemirovsky and Yudin, 1983; Agarwal et al., 2012)
 - **Strongly convex:** $O((\mu n)^{-1})$
Attained by averaged stochastic gradient descent with $\gamma_n \propto (\mu n)^{-1}$
 - **Non-strongly convex:** $O(n^{-1/2})$
Attained by averaged stochastic gradient descent with $\gamma_n \propto n^{-1/2}$
- **Asymptotic analysis of averaging** (Polyak and Juditsky, 1992; Ruppert, 1988)
 - All step sizes $\gamma_n = Cn^{-\alpha}$ with $\alpha \in (1/2, 1)$ lead to $O(n^{-1})$ for **smooth** strongly convex problems

Convex stochastic approximation

Existing work

- **Known global minimax rates of convergence for non-smooth problems** (Nemirovsky and Yudin, 1983; Agarwal et al., 2012)
 - **Strongly convex:** $O((\mu n)^{-1})$
Attained by averaged stochastic gradient descent with $\gamma_n \propto (\mu n)^{-1}$
 - **Non-strongly convex:** $O(n^{-1/2})$
Attained by averaged stochastic gradient descent with $\gamma_n \propto n^{-1/2}$
- **Asymptotic analysis of averaging** (Polyak and Juditsky, 1992; Ruppert, 1988)
 - All step sizes $\gamma_n = Cn^{-\alpha}$ with $\alpha \in (1/2, 1)$ lead to $O(n^{-1})$ for **smooth** strongly convex problems
- **A single algorithm for smooth problems with convergence rate $O(1/n)$ in all situations?**

Least-mean-square algorithm

- **Least-squares:** $f(\theta) = \frac{1}{2}\mathbb{E}[(y_n - \langle \Phi(x_n), \theta \rangle)^2]$ with $\theta \in \mathbb{R}^p$
 - SGD = least-mean-square algorithm (see, e.g., Macchi, 1995)
 - usually studied without averaging and decreasing step-sizes
 - with strong convexity assumption $\mathbb{E}[\Phi(x_n) \otimes \Phi(x_n)] = H \succcurlyeq \mu \cdot \text{Id}$

Least-mean-square algorithm

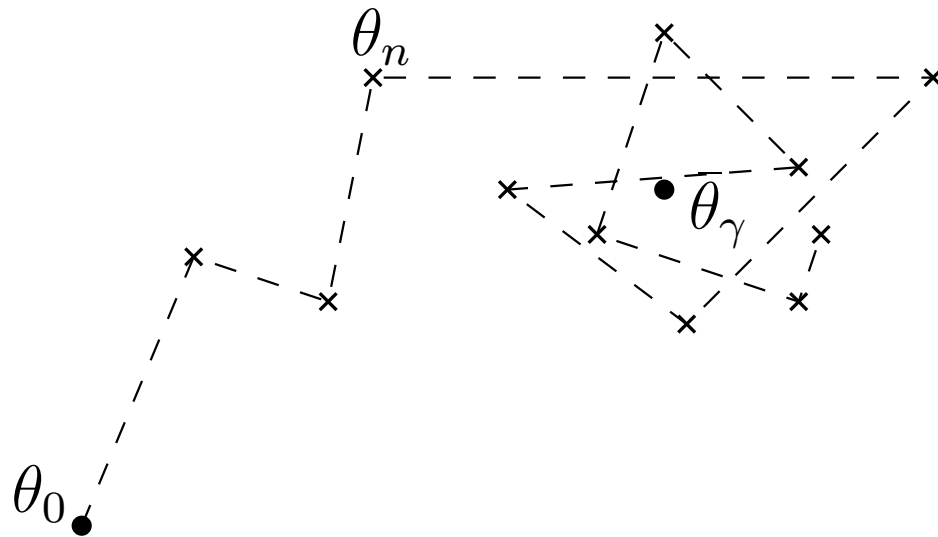
- **Least-squares:** $f(\theta) = \frac{1}{2}\mathbb{E}[(y_n - \langle \Phi(x_n), \theta \rangle)^2]$ with $\theta \in \mathbb{R}^p$
 - SGD = least-mean-square algorithm (see, e.g., Macchi, 1995)
 - usually studied without averaging and decreasing step-sizes
 - with strong convexity assumption $\mathbb{E}[\Phi(x_n) \otimes \Phi(x_n)] = H \succcurlyeq \mu \cdot \text{Id}$
- **New analysis for averaging and constant step-size** $\gamma = 1/(4R^2)$
 - Assume $\|\Phi(x_n)\| \leq R$ and $|y_n - \langle \Phi(x_n), \theta_* \rangle| \leq \sigma$ almost surely
 - **No assumption regarding lowest eigenvalues of H**
 - Main result:
$$\mathbb{E}f(\bar{\theta}_{n-1}) - f(\theta_*) \leq \frac{4\sigma^2 p}{n} + \frac{4R^2 \|\theta_0 - \theta_*\|^2}{n}$$
- **Matches statistical lower bound** (Tsybakov, 2003)
 - Non-asymptotic robust version of Györfi and Walk (1996)

Markov chain interpretation of constant step sizes

- LMS recursion for $f_n(\theta) = \frac{1}{2}(y_n - \langle \Phi(x_n), \theta \rangle)^2$

$$\theta_n = \theta_{n-1} - \gamma(\langle \Phi(x_n), \theta_{n-1} \rangle - y_n)\Phi(x_n)$$

- The sequence $(\theta_n)_n$ is a **homogeneous Markov chain**
 - convergence to a stationary distribution π_γ
 - with expectation $\bar{\theta}_\gamma \stackrel{\text{def}}{=} \int \theta \pi_\gamma(d\theta)$



Markov chain interpretation of constant step sizes

- LMS recursion for $f_n(\theta) = \frac{1}{2}(y_n - \langle \Phi(x_n), \theta \rangle)^2$

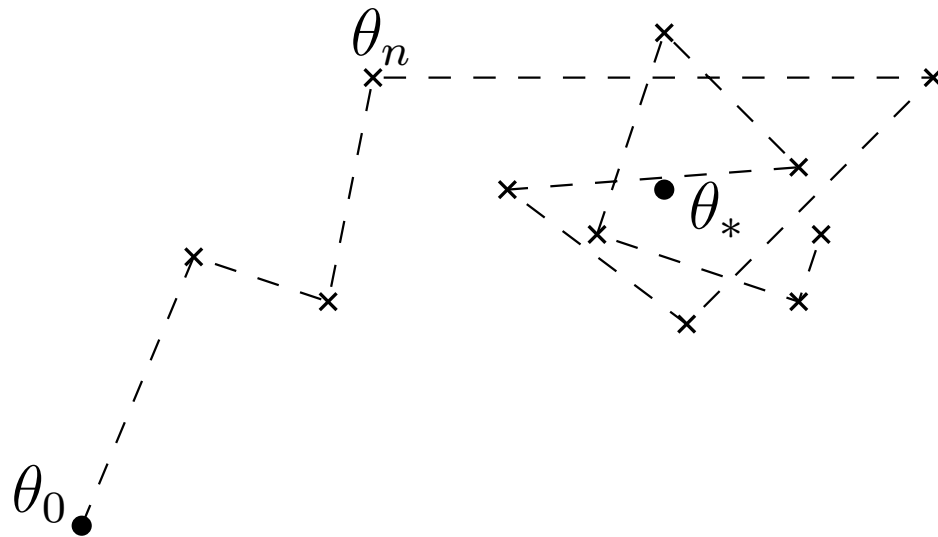
$$\theta_n = \theta_{n-1} - \gamma(\langle \Phi(x_n), \theta_{n-1} \rangle - y_n)\Phi(x_n)$$

- The sequence $(\theta_n)_n$ is a **homogeneous Markov chain**

– convergence to a stationary distribution π_γ

– with expectation $\bar{\theta}_\gamma \stackrel{\text{def}}{=} \int \theta \pi_\gamma(d\theta)$

- **For least-squares, $\bar{\theta}_\gamma = \theta_*$**



Markov chain interpretation of constant step sizes

- LMS recursion for $f_n(\theta) = \frac{1}{2}(y_n - \langle \Phi(x_n), \theta \rangle)^2$

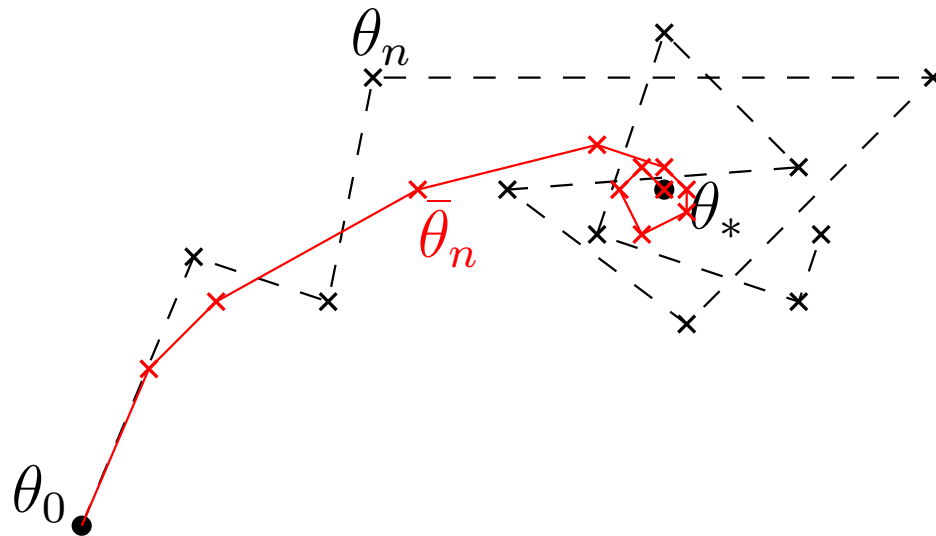
$$\theta_n = \theta_{n-1} - \gamma(\langle \Phi(x_n), \theta_{n-1} \rangle - y_n)\Phi(x_n)$$

- The sequence $(\theta_n)_n$ is a **homogeneous Markov chain**

– convergence to a stationary distribution π_γ

– with expectation $\bar{\theta}_\gamma \stackrel{\text{def}}{=} \int \theta \pi_\gamma(d\theta)$

- **For least-squares, $\bar{\theta}_\gamma = \theta_*$**



Markov chain interpretation of constant step sizes

- LMS recursion for $f_n(\theta) = \frac{1}{2}(y_n - \langle \Phi(x_n), \theta \rangle)^2$

$$\theta_n = \theta_{n-1} - \gamma(\langle \Phi(x_n), \theta_{n-1} \rangle - y_n)\Phi(x_n)$$

- The sequence $(\theta_n)_n$ is a **homogeneous Markov chain**

- convergence to a stationary distribution π_γ

- with expectation $\bar{\theta}_\gamma \stackrel{\text{def}}{=} \int \theta \pi_\gamma(d\theta)$

- **For least-squares, $\bar{\theta}_\gamma = \theta_*$**

- θ_n does not converge to θ_* but oscillates around it

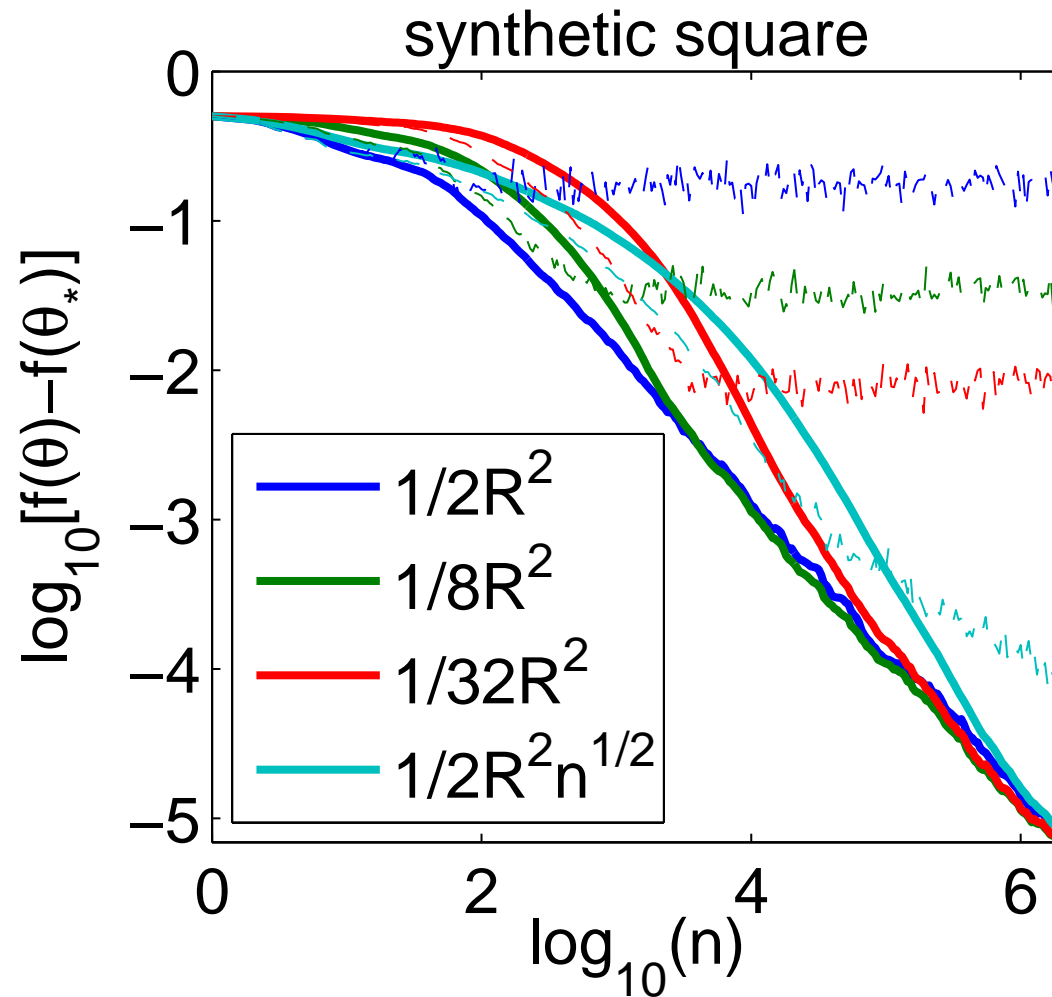
- oscillations of order $\sqrt{\gamma}$

- **Ergodic theorem:**

- Averaged iterates converge to $\bar{\theta}_\gamma = \theta_*$ at rate $O(1/n)$

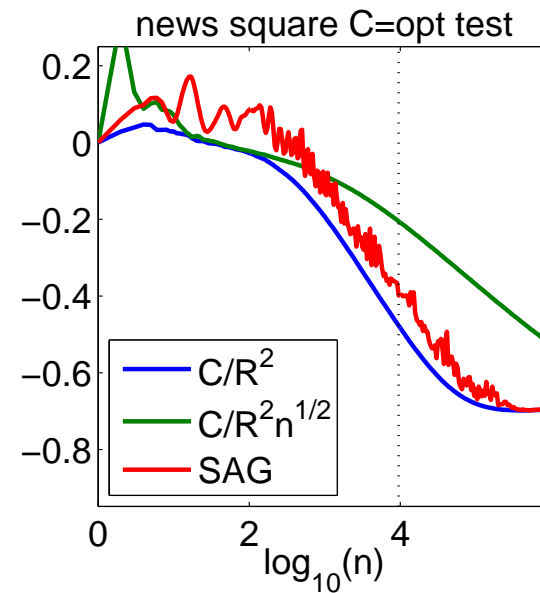
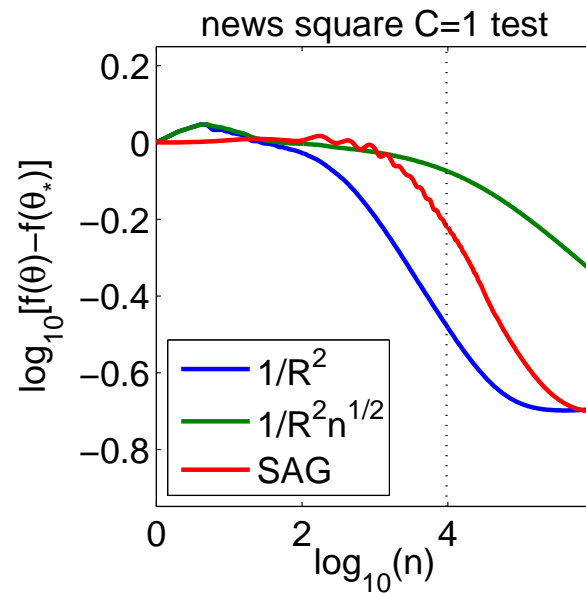
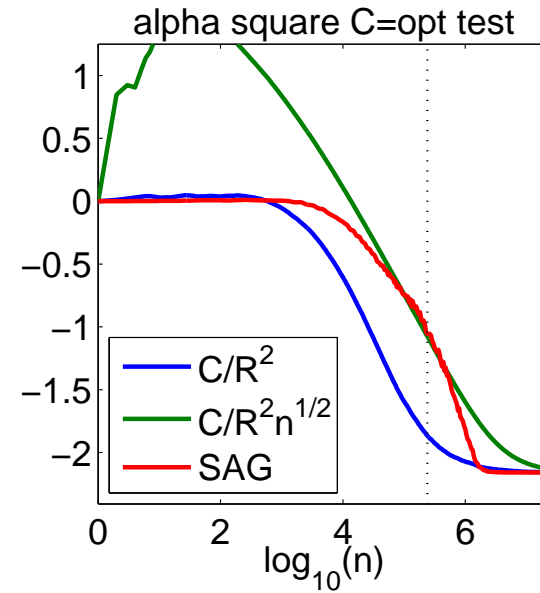
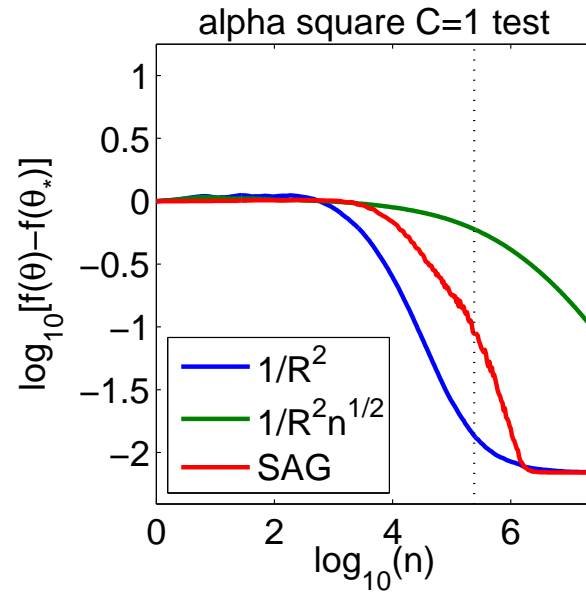
Simulations - synthetic examples

- Gaussian distributions - $p = 20$



Simulations - benchmarks

- *alpha* ($p = 500, n = 500\,000$), *news* ($p = 1\,300\,000, n = 20\,000$)

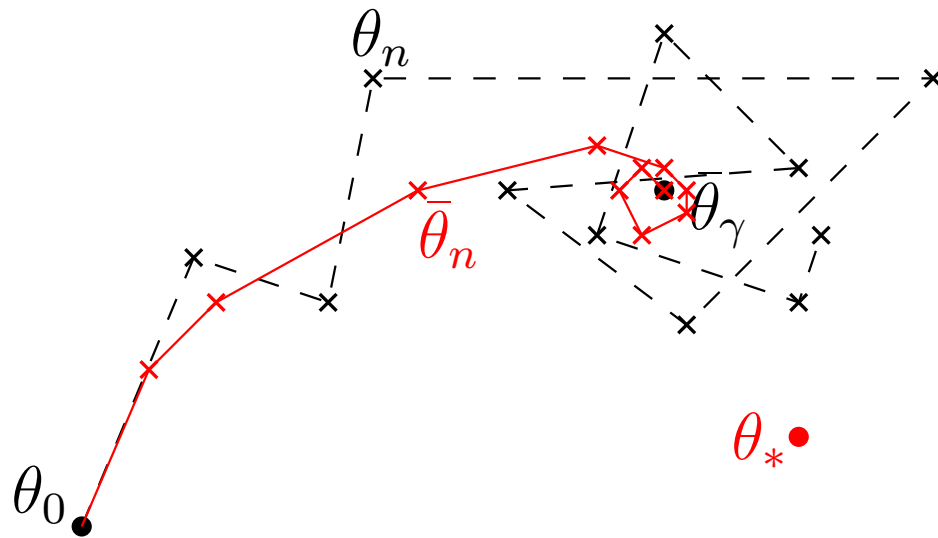


Beyond least-squares - Markov chain interpretation

- Recursion $\theta_n = \theta_{n-1} - \gamma f'_n(\theta_{n-1})$ also defines a Markov chain
 - Stationary distribution π_γ such that $\int f'(\theta)\pi_\gamma(d\theta) = 0$
 - When f' is not linear, $f'(\int \theta\pi_\gamma(d\theta)) \neq \int f'(\theta)\pi_\gamma(d\theta) = 0$

Beyond least-squares - Markov chain interpretation

- Recursion $\theta_n = \theta_{n-1} - \gamma f'_n(\theta_{n-1})$ also defines a Markov chain
 - Stationary distribution π_γ such that $\int f'(\theta)\pi_\gamma(d\theta) = 0$
 - When f' is not linear, $f'(\int \theta\pi_\gamma(d\theta)) \neq \int f'(\theta)\pi_\gamma(d\theta) = 0$
- θ_n oscillates around the wrong value $\bar{\theta}_\gamma \neq \theta_*$

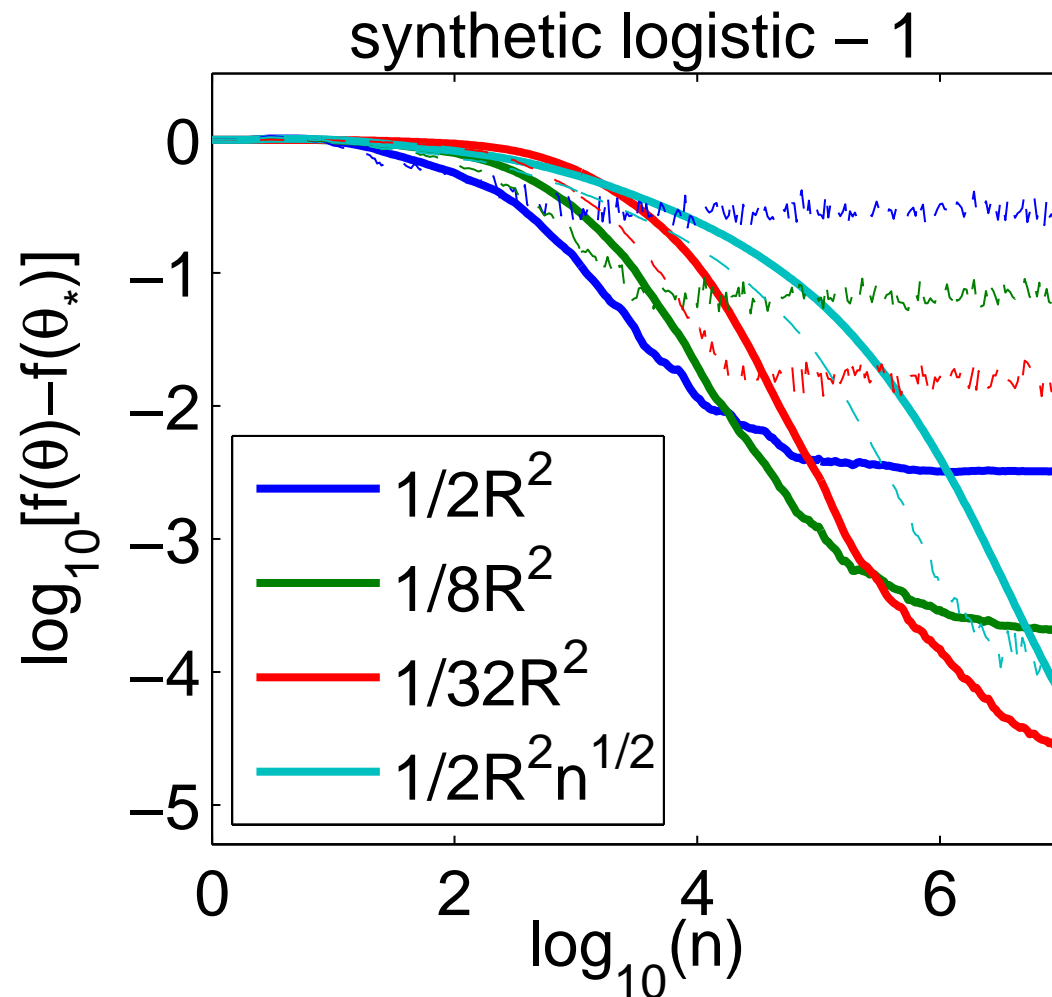


Beyond least-squares - Markov chain interpretation

- Recursion $\theta_n = \theta_{n-1} - \gamma f'_n(\theta_{n-1})$ also defines a Markov chain
 - Stationary distribution π_γ such that $\int f'(\theta)\pi_\gamma(d\theta) = 0$
 - When f' is not linear, $f'(\int \theta\pi_\gamma(d\theta)) \neq \int f'(\theta)\pi_\gamma(d\theta) = 0$
- θ_n oscillates around the wrong value $\bar{\theta}_\gamma \neq \theta_*$
 - moreover, $\|\theta_* - \theta_n\| = O_p(\sqrt{\gamma})$
- Ergodic theorem
 - averaged iterates converge to $\bar{\theta}_\gamma \neq \theta_*$ at rate $O(1/n)$
 - moreover, $\|\theta_* - \bar{\theta}_\gamma\| = O(\gamma)$ (Bach, 2013)

Simulations - synthetic examples

- Gaussian distributions - $p = 20$



Restoring convergence through online Newton steps

- **Known facts**

1. Averaged SGD with $\gamma_n \propto n^{-1/2}$ leads to *robust* rate $O(n^{-1/2})$ for all convex functions
2. Averaged SGD with γ_n constant leads to *robust* rate $O(n^{-1})$ for all convex *quadratic* functions
3. Newton's method squares the error at each iteration for smooth functions
4. A single step of Newton's method is equivalent to minimizing the quadratic Taylor expansion

Restoring convergence through online Newton steps

- **Known facts**

1. Averaged SGD with $\gamma_n \propto n^{-1/2}$ leads to *robust* rate $O(n^{-1/2})$ for all convex functions
2. Averaged SGD with γ_n constant leads to *robust* rate $O(n^{-1})$ for all convex *quadratic* functions $\Rightarrow O(n^{-1})$
3. Newton's method squares the error at each iteration for smooth functions $\Rightarrow O((n^{-1/2})^2)$
4. A single step of Newton's method is equivalent to minimizing the quadratic Taylor expansion

- **Online Newton step**

- Rate: $O((n^{-1/2})^2 + n^{-1}) = O(n^{-1})$
- Complexity: $O(p)$ per iteration

Restoring convergence through online Newton steps

- The Newton step for $f = \mathbb{E}f_n(\theta) \stackrel{\text{def}}{=} \mathbb{E}[\ell(y_n, \langle \theta, \Phi(x_n) \rangle)]$ at $\tilde{\theta}$ is equivalent to minimizing the quadratic approximation

$$\begin{aligned}g(\theta) &= f(\tilde{\theta}) + \langle f'(\tilde{\theta}), \theta - \tilde{\theta} \rangle + \frac{1}{2} \langle \theta - \tilde{\theta}, f''(\tilde{\theta})(\theta - \tilde{\theta}) \rangle \\ &= f(\tilde{\theta}) + \langle \mathbb{E}f'_n(\tilde{\theta}), \theta - \tilde{\theta} \rangle + \frac{1}{2} \langle \theta - \tilde{\theta}, \mathbb{E}f''_n(\tilde{\theta})(\theta - \tilde{\theta}) \rangle \\ &= \mathbb{E} \left[f(\tilde{\theta}) + \langle f'_n(\tilde{\theta}), \theta - \tilde{\theta} \rangle + \frac{1}{2} \langle \theta - \tilde{\theta}, f''_n(\tilde{\theta})(\theta - \tilde{\theta}) \rangle \right]\end{aligned}$$

Restoring convergence through online Newton steps

- The Newton step for $f = \mathbb{E}f_n(\theta) \stackrel{\text{def}}{=} \mathbb{E}[\ell(y_n, \langle \theta, \Phi(x_n) \rangle)]$ at $\tilde{\theta}$ is equivalent to minimizing the quadratic approximation

$$\begin{aligned}g(\theta) &= f(\tilde{\theta}) + \langle f'(\tilde{\theta}), \theta - \tilde{\theta} \rangle + \frac{1}{2} \langle \theta - \tilde{\theta}, f''(\tilde{\theta})(\theta - \tilde{\theta}) \rangle \\ &= f(\tilde{\theta}) + \langle \mathbb{E}f'_n(\tilde{\theta}), \theta - \tilde{\theta} \rangle + \frac{1}{2} \langle \theta - \tilde{\theta}, \mathbb{E}f''_n(\tilde{\theta})(\theta - \tilde{\theta}) \rangle \\ &= \mathbb{E} \left[f(\tilde{\theta}) + \langle f'_n(\tilde{\theta}), \theta - \tilde{\theta} \rangle + \frac{1}{2} \langle \theta - \tilde{\theta}, f''_n(\tilde{\theta})(\theta - \tilde{\theta}) \rangle \right]\end{aligned}$$

- **Complexity of least-mean-square recursion for g is $O(p)$**

$$\theta_n = \theta_{n-1} - \gamma [f'_n(\tilde{\theta}) + f''_n(\tilde{\theta})(\theta_{n-1} - \tilde{\theta})]$$

- $f''_n(\tilde{\theta}) = \ell''(y_n, \langle \tilde{\theta}, \Phi(x_n) \rangle) \Phi(x_n) \otimes \Phi(x_n)$ has rank one
- **New online Newton step without computing/inverting Hessians**

Choice of support point for online Newton step

- **Two-stage procedure**

- (1) Run $n/2$ iterations of averaged SGD to obtain $\tilde{\theta}$
- (2) Run $n/2$ iterations of averaged constant step-size LMS
 - Reminiscent of one-step estimators (see, e.g., Van der Vaart, 2000)
 - **Provable convergence rate of $O(p/n)$** for logistic regression
 - Additional assumptions but no **strong convexity**

Choice of support point for online Newton step

- **Two-stage procedure**

- (1) Run $n/2$ iterations of averaged SGD to obtain $\tilde{\theta}$

- (2) Run $n/2$ iterations of averaged constant step-size LMS

- Reminiscent of one-step estimators (see, e.g., Van der Vaart, 2000)

- **Provable convergence rate of $O(p/n)$** for logistic regression

- Additional assumptions but no **strong convexity**

- **Update at each iteration using the current averaged iterate**

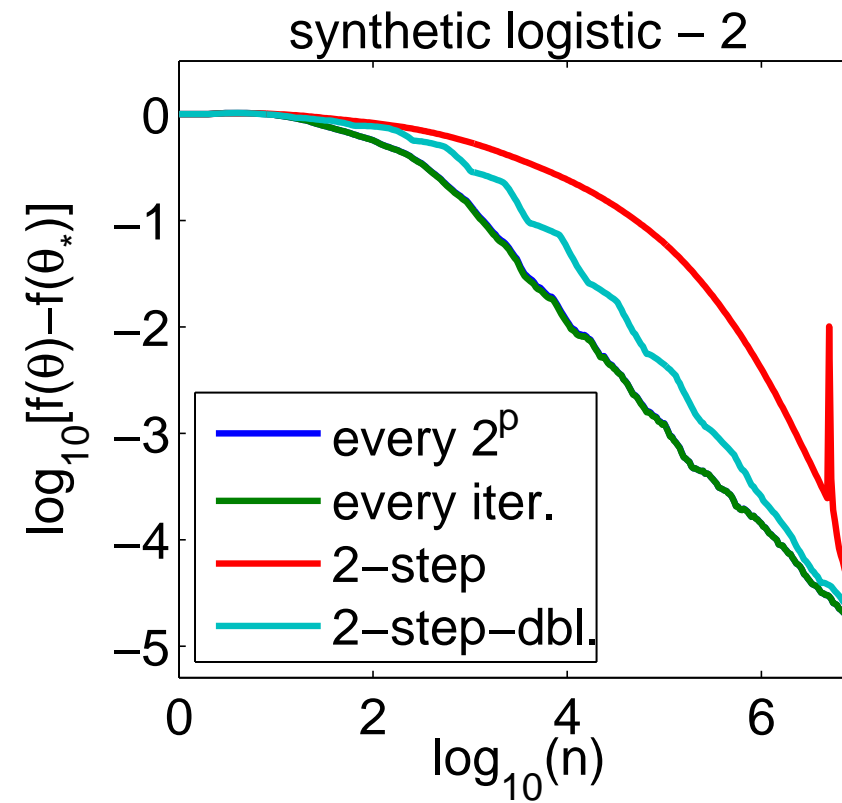
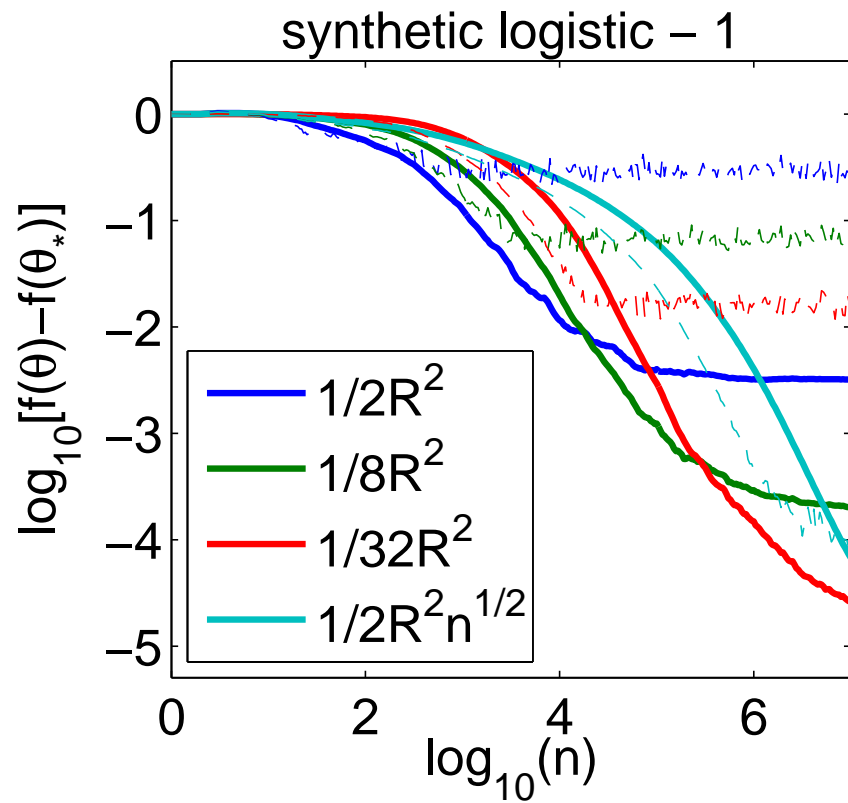
- Recursion:
$$\theta_n = \theta_{n-1} - \gamma [f'_n(\bar{\theta}_{n-1}) + f''_n(\bar{\theta}_{n-1})(\theta_{n-1} - \bar{\theta}_{n-1})]$$

- No provable convergence rate (yet) but best practical behavior

- Note (dis)similarity with regular SGD: $\theta_n = \theta_{n-1} - \gamma f'_n(\theta_{n-1})$

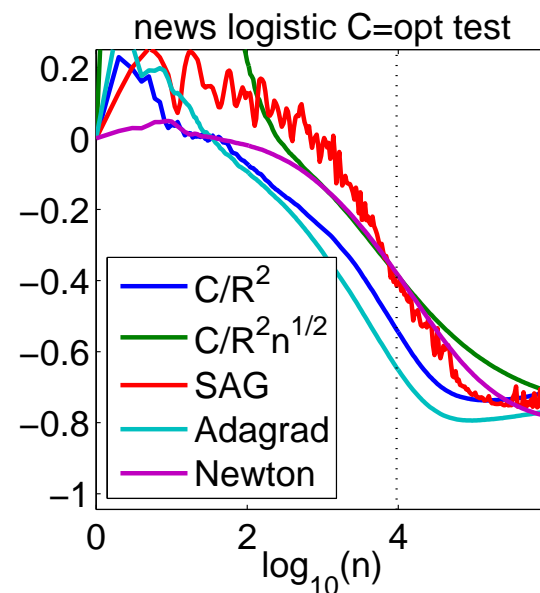
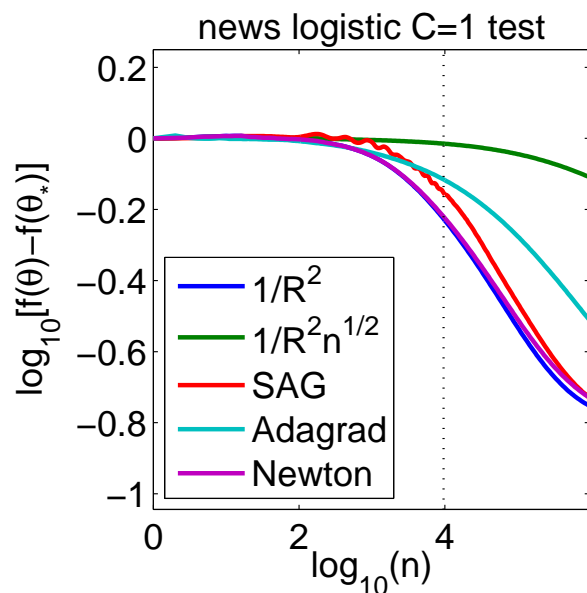
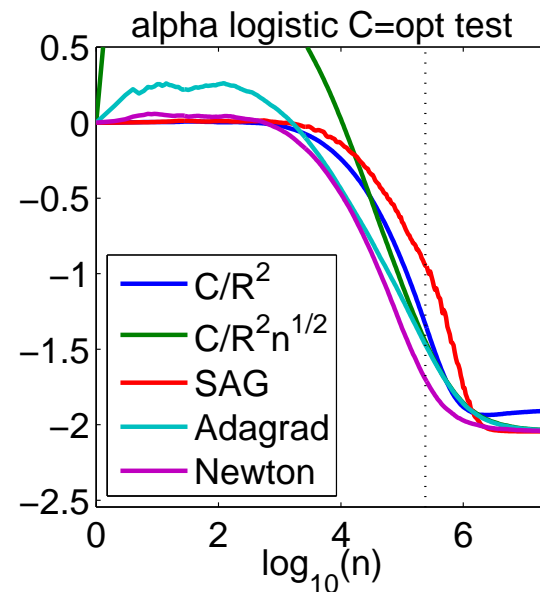
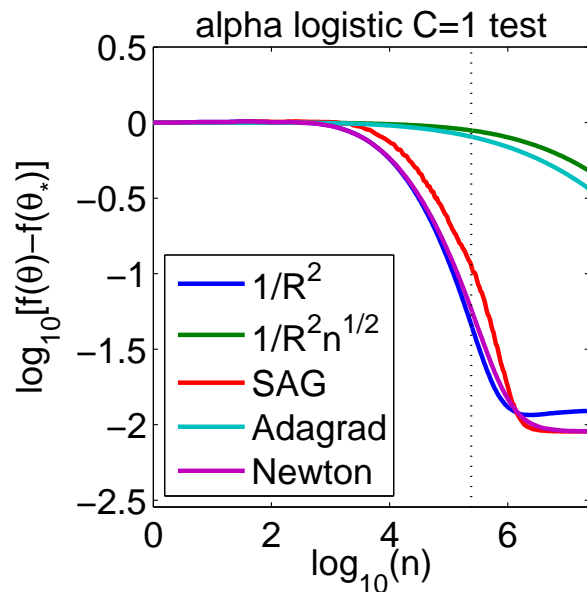
Simulations - synthetic examples

- Gaussian distributions - $p = 20$



Simulations - benchmarks

- *alpha* ($p = 500, n = 500\ 000$), *news* ($p = 1\ 300\ 000, n = 20\ 000$)



Going beyond a single pass over the data

- **Stochastic approximation**

- Assumes infinite data stream
- Observations are used only once
- Directly minimizes **testing** cost $\mathbb{E}_{(x,y)} \ell(y, \langle \theta, \Phi(x) \rangle)$

Going beyond a single pass over the data

- **Stochastic approximation**

- Assumes infinite data stream
- Observations are used only once
- Directly minimizes **testing** cost $\mathbb{E}_{(x,y)} \ell(y, \langle \theta, \Phi(x) \rangle)$

- **Machine learning practice**

- Finite data set $(x_1, y_1, \dots, x_n, y_n)$
- Multiple passes
- Minimizes **training** cost $\frac{1}{n} \sum_{i=1}^n \ell(y_i, \langle \theta, \Phi(x_i) \rangle)$
- Need to regularize (e.g., by the ℓ_2 -norm) to avoid overfitting

- **Goal:** minimize $g(\theta) = \frac{1}{n} \sum_{i=1}^n f_i(\theta)$

Stochastic vs. deterministic methods

- Minimizing $g(\theta) = \frac{1}{n} \sum_{i=1}^n f_i(\theta)$ with $f_i(\theta) = \ell(y_i, \langle \theta, \Phi(x_i) \rangle) + \mu \Omega(\theta)$
- **Batch** gradient descent: $\theta_t = \theta_{t-1} - \gamma_t g'(\theta_{t-1}) = \theta_{t-1} - \frac{\gamma_t}{n} \sum_{i=1}^n f'_i(\theta_{t-1})$
 - Linear (e.g., exponential) convergence rate in $O(e^{-\alpha t})$
 - Iteration complexity is linear in n (*with line search*)

Stochastic vs. deterministic methods

- Minimizing $g(\theta) = \frac{1}{n} \sum_{i=1}^n f_i(\theta)$ with $f_i(\theta) = \ell(y_i, \langle \theta, \Phi(x_i) \rangle) + \mu\Omega(\theta)$
- **Batch** gradient descent: $\theta_t = \theta_{t-1} - \gamma_t g'(\theta_{t-1}) = \theta_{t-1} - \frac{\gamma_t}{n} \sum_{i=1}^n f'_i(\theta_{t-1})$
 - Linear (e.g., exponential) convergence rate in $O(e^{-\alpha t})$
 - Iteration complexity is linear in n (*with line search*)
- **Stochastic** gradient descent: $\theta_t = \theta_{t-1} - \gamma_t f'_{i(t)}(\theta_{t-1})$
 - Sampling with replacement: $i(t)$ random element of $\{1, \dots, n\}$
 - Convergence rate in $O(1/t)$
 - Iteration complexity is independent of n (*step size selection?*)

Stochastic average gradient (Le Roux, Schmidt, and Bach, 2012)

- **Stochastic average gradient (SAG) iteration**
 - Keep in memory the gradients of all functions $f_i, i = 1, \dots, n$
 - Random selection $i(t) \in \{1, \dots, n\}$ with replacement
 - Iteration: $\theta_t = \theta_{t-1} - \frac{\gamma_t}{n} \sum_{i=1}^n y_i^t$ with $y_i^t = \begin{cases} f'_i(\theta_{t-1}) & \text{if } i = i(t) \\ y_i^{t-1} & \text{otherwise} \end{cases}$

Stochastic average gradient (Le Roux, Schmidt, and Bach, 2012)

- **Stochastic average gradient (SAG) iteration**
 - Keep in memory the gradients of all functions $f_i, i = 1, \dots, n$
 - Random selection $i(t) \in \{1, \dots, n\}$ with replacement
 - Iteration: $\theta_t = \theta_{t-1} - \frac{\gamma_t}{n} \sum_{i=1}^n y_i^t$ with $y_i^t = \begin{cases} f'_i(\theta_{t-1}) & \text{if } i = i(t) \\ y_i^{t-1} & \text{otherwise} \end{cases}$
- Stochastic version of incremental average gradient (Blatt et al., 2008)
- Extra memory requirement
 - **Supervised machine learning**
 - If $f_i(\theta) = \ell_i(y_i, \langle \Phi(x_i), \theta \rangle)$, then $f'_i(\theta) = \ell'_i(y_i, \langle \Phi(x_i), \theta \rangle) \Phi(x_i)$
 - Only need to store n real numbers

Stochastic average gradient - Convergence analysis

- **Assumptions**

- Each f_i is L -smooth, $i = 1, \dots, n$
- $g = \frac{1}{n} \sum_{i=1}^n f_i$ is μ -strongly convex (with potentially $\mu = 0$)
- constant step size $\gamma_t = 1/(16L)$
- initialization with one pass of averaged SGD

Stochastic average gradient - Convergence analysis

- **Assumptions**

- Each f_i is L -smooth, $i = 1, \dots, n$
- $g = \frac{1}{n} \sum_{i=1}^n f_i$ is μ -strongly convex (with potentially $\mu = 0$)
- constant step size $\gamma_t = 1/(16L)$
- initialization with one pass of averaged SGD

- **Strongly convex case** (Le Roux et al., 2012, 2013)

$$\mathbb{E}[g(\theta_t) - g(\theta_*)] \leq \left(\frac{8\sigma^2}{n\mu} + \frac{4L\|\theta_0 - \theta_*\|^2}{n} \right) \exp\left(-t \min\left\{\frac{1}{8n}, \frac{\mu}{16L}\right\}\right)$$

- **Linear (exponential) convergence rate with $O(1)$ iteration cost**
- After one pass, reduction of cost by $\exp\left(-\min\left\{\frac{1}{8}, \frac{n\mu}{16L}\right\}\right)$

Stochastic average gradient - Convergence analysis

- **Assumptions**

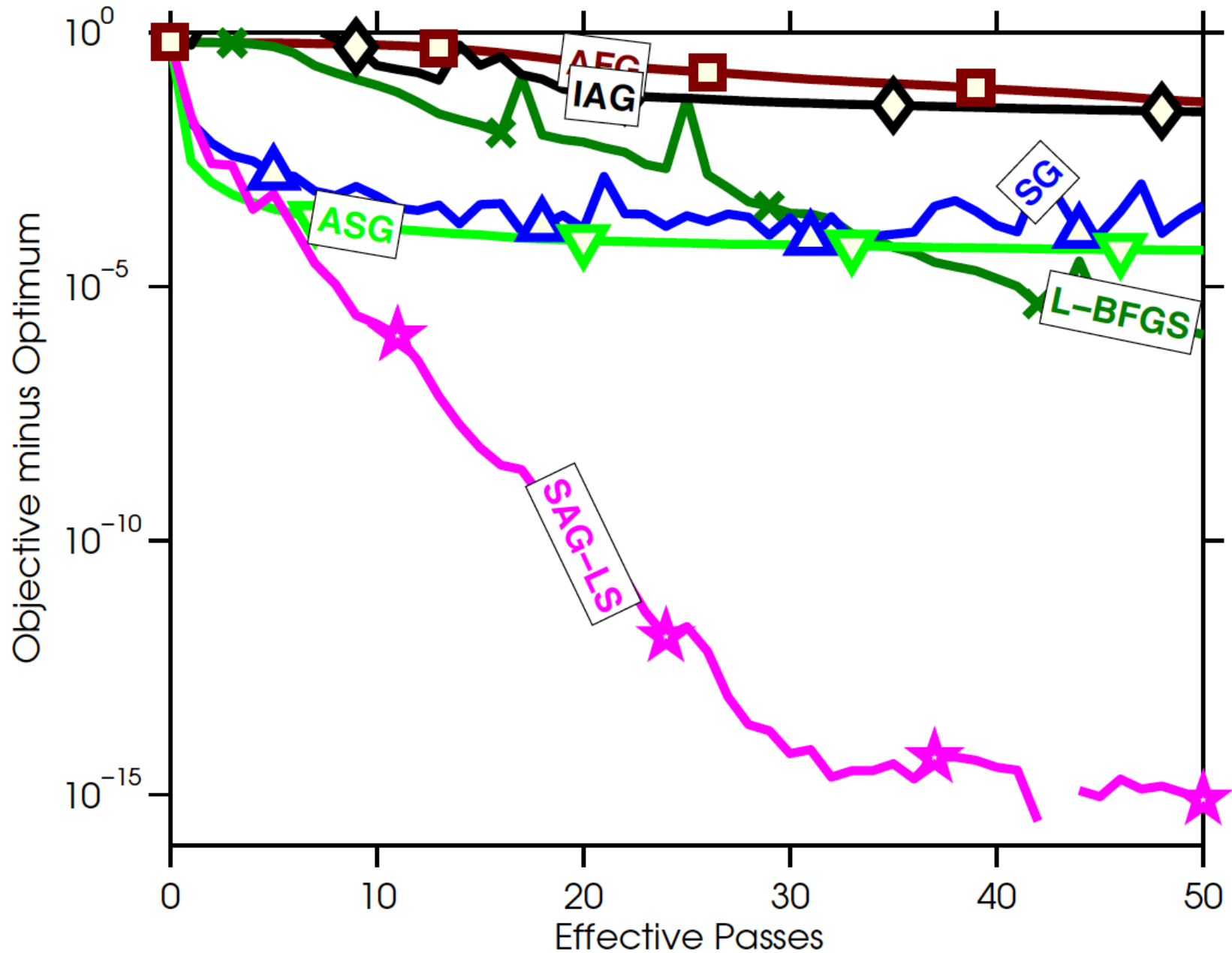
- Each f_i is L -smooth, $i = 1, \dots, n$
- $g = \frac{1}{n} \sum_{i=1}^n f_i$ is μ -strongly convex (with potentially $\mu = 0$)
- constant step size $\gamma_t = 1/(16L)$
- initialization with one pass of averaged SGD

- **Non-strongly convex case** (Le Roux et al., 2013)

$$\mathbb{E}[g(\theta_t) - g(\theta_*)] \leq 48 \frac{\sigma^2 + L \|\theta_0 - \theta_*\|^2}{\sqrt{n}} \frac{n}{t}$$

- Improvement over regular batch and stochastic gradient
- Adaptivity to potentially hidden strong convexity

spam dataset ($n = 92\ 189$, $p = 823\ 470$)



Conclusions

- **Constant-step-size averaged stochastic gradient descent**
 - Reaches convergence rate $O(1/n)$ in all regimes
 - Improves on the $O(1/\sqrt{n})$ lower-bound of non-smooth problems
 - Efficient online Newton step for non-quadratic problems
- **Going beyond a single pass through the data**
 - Keep memory of all gradients for finite training sets
 - Randomization leads to easier analysis **and** faster rates
 - Relationship with Shalev-Shwartz and Zhang (2012); Mairal (2013)
- **Extensions**
 - Non-differentiable terms, **kernels**, line-search, **parallelization**, etc.
 - Beyond supervised learning, beyond convex problems

References

- A. Agarwal, P. L. Bartlett, P. Ravikumar, and M. J. Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *Information Theory, IEEE Transactions on*, 58(5):3235–3249, 2012.
- F. Bach. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. Technical Report 00804431, HAL, 2013.
- F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $o(1/n)$. Technical Report 00831977, HAL, 2013.
- Albert Benveniste, Michel Métivier, and Pierre Priouret. *Adaptive algorithms and stochastic approximations*. Springer Publishing Company, Incorporated, 2012.
- D. Blatt, A. O. Hero, and H. Gauchman. A convergent incremental gradient method with a constant step size. *SIAM Journal on Optimization*, 18(1):29–51, 2008.
- L. Bottou and O. Bousquet. The tradeoffs of large scale learning. In *Adv. NIPS*, 2008.
- L. Györfi and H. Walk. On the averaged stochastic approximation for linear regression. *SIAM Journal on Control and Optimization*, 34(1):31–61, 1996.
- N. Le Roux, M. Schmidt, and F. Bach. A stochastic gradient method with an exponential convergence rate for strongly-convex optimization with finite training sets. In *Adv. NIPS*, 2012.
- N. Le Roux, M. Schmidt, and F. Bach. A stochastic gradient method with an exponential convergence rate for strongly-convex optimization with finite training sets. Technical Report 00674995, HAL, 2013.

- O. Macchi. *Adaptive processing: The least mean squares approach with applications in transmission*. Wiley West Sussex, 1995.
- Julien Mairal. Optimization with first-order surrogate functions. *arXiv preprint arXiv:1305.3120*, 2013.
- A. S. Nemirovsky and D. B. Yudin. *Problem complexity and method efficiency in optimization*. Wiley & Sons, 1983.
- B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.
- H. Robbins and S. Monro. A stochastic approximation method. *Ann. Math. Statistics*, 22:400–407, 1951. ISSN 0003-4851.
- D. Ruppert. Efficient estimations from a slowly convergent Robbins-Monro process. Technical Report 781, Cornell University Operations Research and Industrial Engineering, 1988.
- S. Shalev-Shwartz and T. Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. Technical Report 1209.1873, Arxiv, 2012.
- A. B. Tsybakov. Optimal rates of aggregation. In *Proc. COLT*, 2003.
- A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge Univ. press, 2000.